Introducing the UraLUID Database

Uralic **L**anguages **U**nder the **I**nfluence **D**atabase

Eszter Simon – Ágnes Kalivoda

Budapest, 30 November 2017

Research Institute for Linguistics, Hungarian Academy of Sciences

The Outline of the Presentation

1. Introduction
2. The Description of the Database
3. Text and Speech Processing
4. The Structure of the Database
5. How to Use the Database

Introduction

About the Project

Languages under the Influence.
Uralic syntax changing in an asymmetrical contact situation

- Research Institute for Linguistics, Hungarian Academy of Sciences
- January 2016 – July 2017
- supported by the National Research Development and Innovation Office (ERC_HU_15 118079)
- project investigator: Katalin É. Kiss
- interdisciplinary team: depts of Finno-Ugric and Historical Linguistics, Language Technology, and Theoretical Linguistics

Aims of the Project

Theoretical work

syntactic changes affected by the influence of Russian in Uralic languages

Computational work

linguistically annotated database

Languages

Synya Khanty, Surgut Khanty, Udmurt, and Tundra Nenets

The Criteria of Database Building

- systematic: creating a systematically annotated database, not an eclectic data collection
- standard: following international standards → Unicode, IPA, ELAN
- primary data: documenting the languages in their natural forms, as close to the everyday use of the language as possible → spontaneous speech, interviews, and blog posts
- representative: balancing sociolinguistical parameters, dialects, and genres → written and spoken data from several sources and diverse genres
- multipurpose: making data available for further research
- open: making data freely available

The Description of the Database

The Date of Source Data

in order to observe syntactic changes,
we collected old and new texts

	Synya Khanty	Surgut Khanty	Udmurt	Tundra Nenets
old			1885	
			1893	
		1901		1911-12
	1936-37			
new				1937-1980
				1995
		2000, 2004		1998-2011
	2011	2017	2013-14	2017

The Genre and Modality of the Data

		Synya Kh.	Surgut Kh.	Udmurt	T. Nenets
old	written	folklore	folklore	folklore	folklore
	spoken		folklore		
new	written		interview	blog	folklore newspaper
	spoken	narrative expository	narrative expository		narrative expository

Volume of Data

Planned:

4000–4000 tokens of old and new texts for each language

Now:

		Synya Kh.	Surgut Kh.	Udmurt	T. Nenets
old	written	6,974	3,758	4,212	5,057
	spoken		00:36:17		
new	written		3,188	4,467	81,190
	spoken	01:08:22	00:28:58		01:57:40

Text and Speech Processing

Text Processing Steps

1. scan and OCR → original text
2. character-level normalization → UTF-8 encoded plain text files
3. transcription and transliteration → converted texts into FUT, IPA, Cyrillic
4. morphological annotation → lemma, POS tags, inflectional codes, English and Hungarian glosses
5. translation → English, Hungarian, German, Russian translations
6. segmentation → text segmented into sentences, sentences segmented into tokens, tokens segmented into morphemes
7. alignment → glosses aligned on morpheme- or token level, lemma and POS aligned on token level, translations aligned on sentence level

Speech Processing

1. transcription → FUT or IPA
2. time-alignment → sentences aligned to the time slots of the audio file
3. segmentation → sentence-level segmented transcription → sentences segmented into tokens, tokens segmented into morphemes
4. adding external annotation → morphological annotation and translations
5. alignment → time-aligned sentences, glosses aligned on morpheme- or token level, lemma and POS aligned on token level, translations aligned on sentence level

The Structure of the Database

The Structure of the Database

http://www.nytud.hu/depts/tlp/uralic/dbases.html

1. source
2. metadata (detailed information about the given text)
3. information on morphological analysis
4. text (split into sentences, with different transcriptions)
5. morphologically analyzed text
6. translations
7. ELAN files (every annotation level united in an `.eaf` file + audio)

Metadata

- identification
  - title
  - text ID
  - page number
  - file name
- genre
- measurement
  - token number
  - duration
- (sub)dialect
- info about the informant
  - informant's age
  - informant's gender
- stimulus

Text

- the original text is split into sentences
- transcriptions and transliterations:
  - original transcription
  - FU transcription(s)
  - Cyrillic
  - **IPA** (obligatory)

Morphologically Analyzed text

- stored as `.tsv` files with 15 fixed columns containing all token-level information
  - one token per line
  - sentence boundaries are marked by empty lines
- the structure of the columns:
  - 1-9: the token and its transcriptions
  - 10: segmented token
  - 11: lemma
  - 12: Hungarian gloss
  - 13: English gloss
  - 14: POS tag (and semantic label if there is any)
  - 15: RUS label (if the word has Russian origin)

Translations

- **English** (obligatory)
- Russian
- German
- Hungarian

all translations, transcriptions and transliterations are sentence-level aligned with the original text

ELAN files

- `.eaf` file: containing all data aligned on sentence-, token- and morpheme-level
- audio file: `wav` or `wma`
- the original sentences are aligned to the time slots of the audio file
- the other pieces of information are connected to the sentences via symbolic references

How to Use the Database

How to Use the Database

1. use the `.eaf` files with the corresponding audio files:
   - download the latest version of ELAN: `https://tla.mpi.nl/tools/tla-tools/elan/`
   - download the Charis SIL font package: `https://software.sil.org/charis/`
   - download the `.eaf` and audio files: `http://www.nytud.hu/depts/tlp/uralic/dbases.html`
   - open ELAN → Open... → choose the needed `.eaf` file and the corresponding audio file
2. use the `.tsv` files
   - download the `.tsv` files
   - use Unix commands or your own statistical tools

Acknowledgements

Our thanks
for their contribution to the building of the database
goes to:

Asztalos, Erika
Csepregi, Márta
Fejes, László
Gugán, Katalin
Kalmár, Éva
Khanina, Olesya
Kozlov, Aleksey
Longortov, Arkady Petrovich
Mus, Nikolett
Németh, Szilvia
Nguyen-Dang, Nóra Lien
Okotetto, Khadry
Pesikova, Agrafena
Ruttkay-Miklián, Eszter
Schön, Zsófia
Sipos, Mária
Skribnik, Elena
Speshilova, Yulia
Tánczos, Orsolya
Volkova, Anisya

Thank you for your attention!

simon.eszter@nytud.mta.hu
kalivoda.agnes@nytud.mta.hu

http://www.nytud.hu/depts/tlp/uralic/dbases.html